

МОЖЕТ ЛИ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ СОЗДАВАТЬ СИМУЛЯЦИОННЫЕ СЦЕНАРИИ НАРАВНЕ С ЧЕЛОВЕКОМ? СРАВНИТЕЛЬНЫЙ АНАЛИЗ СЦЕНАРИЕВ, НАПИСАННЫХ ЧЕЛОВЕКОМ И ИИ, НА ПРИМЕРЕ КЕЙСА «ПОЖАР В ОПЕРАЦИОННОЙ»

Горшков Максим Дмитриевич^{1, 2}, Андреев Александр Александрович³, Буланов Роман Леонидович⁴, Долгина Ирина Ивановна⁵, Голубева Олеся Валентиновна⁶, Грибков Денис Михайлович⁷, Зарипова Зульфия Абдулловна⁸, Косцова Надежда Григорьевна⁹, Непершина Валия Рефатовна¹⁰, Ходус Сергей Васильевич¹¹, Шубина Любовь Борисовна⁷

¹ Общероссийская общественная организация «Российское общество симуляционного обучения в медицине» / РОСОМЕД, г. Москва, Российская Федерация

² ЕвроМедСим, Европейский институт симуляции в медицине г. Штуттгарт, Федеративная Республика Германия

³ Городская Мариинская больница, г. Санкт-Петербург, Российская Федерация

⁴ Северный государственный медицинский университет, г. Архангельск, Российская Федерация

⁵ Курский государственный медицинский университет, г. Курск, Российская Федерация

⁶ Национальный медицинский исследовательский центр «Межотраслевой научно-технический комплекс „Микрохирургия глаза“ им. академика С. Н. Фёдорова», г. Москва, Российская Федерация

⁷ Факультет фундаментальной медицины Московского государственного университета им. М. В. Ломоносова, г. Москва, Российская Федерация

⁸ Первый Санкт-Петербургский государственный медицинский университет им. академика И. П. Павлова, г. Санкт-Петербург, Российская Федерация

⁹ Российский университет дружбы народов, г. Москва, Российская Федерация

¹⁰ Медицинский колледж, г. Москва, Российская Федерация

¹¹ Амурская государственная медицинская академия, г. Благовещенск, Российская Федерация

ORCID: Горшков М. Д. 0000-0003-0446-0787

ORCID: Андреев А. А. 0000-0002-5542-9280

ORCID: Буланов Р. Л. 0000-0003-4514-9717

SPIN: Долгина И. И. 8024-0375

SPIN: Голубева О. В. 8711-6297

ORCID: Зарипова З. А. 0000-0002-2224-7536

ORCID: Косцова Н. Г. 0000-0001-9708-9643

SPIN: Непершина В. Р. 9066-3903

ORCID: Ходус С. В. 0000-0001-5138-3791

ORCID: Шубина Л. Б. 0000-0002-4589-5712

gorshkov@rosomed.ru

DOI: 10.46594/2687-0037_2025_4_2151

Аннотация. Слепое сравнительное пилотное исследование посвящено оценке качества симуляционных сценариев, созданных генеративными моделями искусственного интеллекта (ИИ) и человеком. Глобальной целью исследования была комплексная оценка возможности создания сценариев моделями ИИ. Отсюда следовали задачи сопоставления качества ИИ-сценариев с человеческими и оценка способности экспертов определять авторство. Три модели (Grok, ChatGPT, DeepSeek) генерировали сценарии по единому стандартизированному промпту; референтным эталоном служил конкурсный кейс РОСОМЕД. Пять независимых экспертов оценивали четыре сценария по оригинальной шкале ШОСС (8 критериев, 4-балльная система). Лучшим оказался сценарий ChatGPT (средний балл 3, 4), превосходящий человеческий сценарий (2,9). ИИ-оценщик в целом ранжировал сценарии сходно с людьми-экспертами, но при самооценке проявилась его предвзятость по отношению к собственной работе.

Ключевые слова: искусственный интеллект, генеративный ИИ, LLM, симуляционное обучение, клинические сценарии, оценка качества, медицинское образование, пожар в операционной.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Для цитирования: Горшков М. Д., Андреев А. А., Буланов Р. Л., Долгина И. И., Голубева О. В., Грибков Д. М., Зарипова З. А., Косцова Н. Г., Непершина В. Р., Ходус С. В., Шубина Л. Б. Может ли ИИ создавать симуляционные сценарии наравне с человеком? Сравнительный анализ сценариев, написанных человеком и ИИ, на примере кейса «Пожар в операционной» // Виртуальные технологии в медицине. 2025. № 4. С. 362–369. DOI: 10.46594/2687-0037_2025_4_2151

Научная специальность: 3.2.3. Общественное здоровье и организация здравоохранения, социология и история медицины.

Поступила в редакцию 06 декабря 2025 г.

Поступила после рецензирования 12 января 2026 г.

Принята к публикации 12 января 2026 г.

CAN ARTIFICIAL INTELLIGENCE CREATE SIMULATION SCENARIOS ON PAR WITH HUMANS? A COMPARATIVE ANALYSIS OF SCENARIOS WRITTEN BY HUMANS AND AI, USING THE CASE OF «FIRE IN THE OPERATING ROOM»

Gorshkov Maxim^{1,2}, Andreenko Alexander³, Bulanov Roman⁴, Dolgina Irina⁵, Golubeva Olesya⁶,
Gribkov Denis⁷, Zaripova Zulfia⁸, Kostsova Nadezhda⁹, Nepershina Valia¹⁰,
Khodus Sergey¹¹, Shubina Lyubov⁷

¹ All-Russian Public Organization «Russian Society for Simulation Education in Medicine» / ROSOMED, Moscow, Russian Federation

² EuroMedSim, European Institute of Simulation in Medicine, Stuttgart, Federal Republic of Germany

³ Mariinsky City Hospital, St. Petersburg, Russian Federation

⁴ Northern State Medical University, Arkhangelsk, Russian Federation

⁵ Kursk State Medical University, Kursk, Russian Federation

⁶ S. Fyodorov Eye Microsurgery Federal State Institution, Moscow, Russian Federation

⁷ Faculty of Fundamental Medicine, Lomonosov Moscow State University, Moscow, Russian Federation

⁸ First Pavlov State Medical University of St. Petersburg, St. Petersburg, Russian Federation

⁹ RUDN, Moscow, Russian Federation

¹⁰ Medical College, Moscow, Russian Federation

¹¹ Amur State Medical Academy, Blagoveshchensk, Russian Federation

gorshkov@rosomed.ru

DOI: 10.46594/2687-0037_2025_4_2151

Annotation. A blind comparative pilot study was conducted to evaluate the quality of simulation scenarios generated by generative artificial intelligence (AI) models compared to human-generated scenarios. The objectives of the study were to test the feasibility of generating scenarios by AI models, compare their quality with human-generated scenarios, and assess the ability of experts to determine authorship. Three models (Grok, ChatGPT and DeepSeek) were used to generate scenarios based on a standardized prompt, and the ROSOMED competition case served as the reference standard. Five independent experts evaluated four scenarios using the original scale of assessment of a simulation scenario (SASS). The ChatGPT scenario (average score of 3.4) outperformed the human scenario (2.9). The AI evaluator generally ranked the scenarios similarly to the human experts, but it showed bias towards its own work when self-evaluating.

Keywords: artificial intelligence, generative AI, LLM, simulation education, clinical scenarios.

Conflict of interest. The authors declare no conflict of interest.

For quotation: Gorshkov M., Andreenko A., Bulanov R., Dolgina I., Golubeva O., Gribkov D., Zaripova Z., Kostsova N., Nepershina V., Khodus S., Shubina L. Can AI Create Simulation Scenarios on Par with Humans? A Comparative Analysis of Scenarios Written by Humans and AI, Using the Case of “Fire in the Operating Room” // Virtual Technologies in Medicine. 2025. No. 4. P. 362–369. DOI: 10.46594/2687-0037_2025_4_2151

Received December 06, 2025

Revised January 12, 2026

Accepted January 12, 2026

Введение

Симуляционное обучение прочно вошло в современное медицинское образование, став золотым стандартом для отработки практических навыков и командной работы в безопасных условиях [2]. Качество симуляционного занятия в значительной степени определяется глубиной и точностью проработки клинического сценария, на основе которого организуется и проводится тренинг. Разработка качественных сценариев является трудоемким, порой многодневным процессом, требующим от экспертов глубоких предметных знаний, понимания педагогических принципов и клинической практики [1].

Широкое распространение генеративных моделей ИИ в медицинском образовании поднимает ряд вопросов. Ключевой из них: способны ли эти технологии взять на себя или, по крайней мере, существенно облегчить разработку симуляционных сценариев? Предварительные исследования показывают, что ИИ

способен успешно генерировать тексты, планы лекций и тестовые задания [3]. Однако применительно к сложным, комплексным симуляционным сценариям, требующим интеграции клинических, педагогических и организационных аспектов, эта тема остается малоизученной.

Цели и задачи

Целью исследования является комплексная оценка возможности создания симуляционных клинических сценариев генеративными моделями искусственного интеллекта.

Задачи исследования сформулированы следующим образом:

1. Оценить принципиальную возможность создания симуляционных сценариев системами генеративного ИИ.

2. Сравнить качество сценариев, разработанных системами ИИ и людьми.
3. Оценить способность людей-экспертов определить авторство сценариев (ИИ или человек).
4. Оценить способность ИИ-модели определять авторство сценариев (ИИ или человек).

Дизайн исследования

Слепое сравнительное пилотное исследование.

Материалы и методы

На конференции Общероссийской общественной организации «Российское общество симуляционного обучения в медицине» (РОСОМЕД) с 2023 г. проводится конкурс клинических симуляционных сценариев «Лаборатория симуляции», организаторами которого являются Д. М. Грибков, Л. Б. Шубина. Присланные на конкурс кейсы оцениваются по предложенной организаторами оригинальной таблице структурной оценки, содержащей три варианта ответа по четырем критериям. Среди присланных на РОСОМЕД-2025 работ в числе наиболее интересных кейсов рецензентами и организаторами конкурса был отмечен сценарий «Профилактика и управление пожаром в операционной» (автор В. Р. Непершина; рецензенты, дополнившие и переработавшие сценарий, Н. Г. Косцова, О. В. Голубева), который и был взят в качестве референтного для данного исследования.

Для генерации симуляционных сценариев были отобраны три платформы генеративного искусственного интеллекта (ИИ), так называемые Большие языковые модели (Large Language Models, LLM):

- **Grok**, компании X.AI Corp. (xAI), г. Пало Альто, Калифорния, США [<https://grok.com/>];
- **ChatGPT**, компании OpenAI Inc., г. Сан Франциско, Калифорния, США [<https://chatgpt.com/>];
- **DeepSeek**, компании Hangzhou DeepSeek Artificial Intelligence Co., Ltd., г. Ханчжоу, Чжэцзян, Китай [<https://chat.deepseek.com/>].

Выбор моделей основывался на предпочтениях, сформированных текущей доступностью и практикой использования ведущим автором. В целях стандартизации создаваемых сценариев ведущим автором был составлен промпт из 259 слов [Врезка № 1], содержащий название темы, целевую аудиторию, предполагаемый объем, перечень учебных целей (УЦ) и структуру сценария, — т. е. распоряжение всем трем системам ИИ давалось в единой форме и содержало конкретные, специфические указания, направленные, главным образом, на придание однообразия с исходным референтным сценарием. Ключевыми задачами являлись следующие:

- 1) Создать учебный симуляционный сценарий по теме «Пожар в операционной»;
- 2) Привести два реалистичных варианта причин возгорания в операционной;
- 3) Придерживаться Рекомендаций о предотвращении и эффективном управлении пожаром в операционной Американского общества анестезиологов;
- 4) Приводить примеры из Российской и зарубежной практики;
- 5) Использовать в сценарии симулятор пациента;
- 6) Предложить варианты проведения дебрифинга.

Промпт для генерации симуляционных сценариев

Составь клинический кейс по «Профилактике и управлению пожаром в операционной» для операционных сестер и сестер-анестезистов.

Разработай подробный документ примерно на 15–20 тысяч знаков для проведения клинического симуляционного сценария согласно загруженному промпту. Выполни его в формате инструктажа для сотрудников симуляционного центра для внутреннего регламента. Подробно остановись на действиях обучаемых. Уточни хронометраж не только симуляции, но и всего занятия с учетом двух сценариев.

Цели обучения, основная идея сценария:

1. Выявление ситуации, способствующей возникновению пожара.
2. Определение элементов протокола реагирования на пожар.
3. Управление ресурсами в условиях кризиса.

За основу возьми рекомендации о предотвращении и эффективном управлении пожаром в операционной ASA (American Society of Anesthesiologists).

В сценарии должны быть разделы:

- «Актуальность». — Раскрываются важность проблемы, статистика, примеры из российской и зарубежной практики, факторы опасности пожара именно в операционной, психологические и материальные последствия.
- «Брифинг». — Вводная информация для командного тренинга должна быть предоставлена так, чтобы участники симуляционного занятия не узнали о предстоящей имитации пожара. Им следует озвучить учебные цели по командной работе, командной коммуникации и управлению ресурсами в кризисе. Контракт Лас-Вегаса, Контракт на реализм.
- «Ход сценария». — Опиши ход сценария, разбив его по частям. Придумай два реалистичных варианта сценария возгорания. Опиши необходимые действия участников, их роли. Завершением сценария должно

быть восстановление проходимости дыхательных путей пациента. Распиши симуляционное занятие по минутам.

- «Оснащение». — Опиши обстановку, медицинское оборудование и документацию для оснащения помещения симуляции. Опиши симуляционное оборудование и его характеристики (используй симулятор пациента в качестве оперируемого больного).
- «Дебрифинг». — Сформулируй вопросы для дебрифинга, на что следует обратить внимание дебриферу, в частности механизмы возникновения пожара, меры предотвращения, реагирования. Предложи модель, по которой можно проводить дебрифинг.
- «Заключение».

Все кейсы были сгенерированы в один день, 4 ноября 2025 г., и были скопированы из браузера в текстовый файл формата *.doc, что обезличило авторство. Из текстов, сгенерированных ИИ, были удалены немногочисленные очевидные признаки, которые могли указать на их искусственное происхождение. Исходный, референтный сценарий содержал блок-схему действий при пожаре, поэтому для придания единообразия всем оцениваемым документам блок-схема была преобразована в текстовый фрагмент. Другие стилистические или грамматические правки ни одного из сценариев не производились. Все четыре документа были названы так, чтобы исключить косвенные подсказки, указывающие на авторство.

В целях объективной структурированной оценки сценариев ведущим автором была расширена система

оценки, использованная в ходе «Лаборатории симуляции» (Грибков Д. М., Шубина Л. Б., 2024–2025), и создана оригинальная Шкала оценки симуляционных сценариев (ШОСС). В данную шкалу были включены 8 критериев с оценкой по 4-балльной шкале от 1 («Ниже ожиданий») до 4 («Превышает ожидания»). Оценивались следующие критерии: соответствие учебных целей (УЦ) уровню обучаемых; четкие, конкретные, измеримые УЦ; описание приобретаемых навыков; реалистичность сценария; соответствие навыков нормативам; соответствие оснащения нормативам; выполнимость, реализуемость сценария; необычность, оригинальность, проблемность сценария. Для удобства экспертов каждая ячейка содержала краткое описание данной оценки (подробнее см. табл. 1).

Таблица 1

Шкала оценки симуляционного сценария, ШОСС

№	Ниже ожиданий	Близко к ожиданиям	Соответствует ожиданиям	Превышает ожидания
1.	1.1. Учебные цели не соответствуют уровню навыков и знаний целевой аудитории	1.2. Учебные цели частично соответствуют уровню навыков и знаний целевой аудитории	1.3. Учебные цели в целом соответствуют уровню навыков и знаний целевой аудитории	1.4. Полностью соответствуют и достижимы для уровня навыков и знаний целевой аудитории
2.	2.1. Учебные цели не обозначены конкретными, измеримыми, ориентированными на практику действиями, актуальными или ограниченными по времени	2.2. Учебные цели редко или плохо обозначены конкретными, измеримыми, ориентированными на практику действиями, актуальными или ограниченными по времени	2.3. Учебные цели обозначены конкретными, измеримыми, ориентированными на практику действиями, актуальными или ограниченными по времени	2.4. Все учебные цели конкретны, измеримы, ориентированы на практику, актуальны и ограничены по времени
3.	3.1. В сценарии не описаны приобретаемые в ходе симуляции знания и навыки	3.2. В сценарии нечетко и неконкретно описаны приобретаемые в ходе симуляции знания и навыки	3.3. В сценарии описаны приобретаемые в ходе симуляции знания и навыки	3.4. В сценарии подробно, обстоятельно и четко прописаны приобретаемые в ходе симуляции знания и навыки
4.	4.1. Описанный в симуляции случай нереальный, в клинической практике так не бывает	4.2. Описанный в симуляции случай возможен, но действия описаны неправильно, нереалистично, требуется доработка	4.3. Описанный в симуляции случай реален и может случиться в практике	4.4. Описанный в симуляции случай реален, случается в практике и требует тренинга и отработки
5.	5.1. Целевые действия обучаемых (манипуляции, решения и т. п.) абсолютно не соответствуют нормативной базе и клинической практике	5.2. Целевые действия обучаемых (манипуляции, решения и т. п.) частично соответствуют нормативной базе и клинической практике	5.3. Целевые действия обучаемых (манипуляции, решения и т. п.) соответствуют нормативной базе и клинической практике	5.4. Целевые действия обучаемых (манипуляции, решения и т. п.) полностью соответствуют нормативной базе и клинической практике
6.	6.1. Предлагаемое оснащение абсолютно не соответствует современным нормативам	6.2. Предлагаемое оснащение отчасти соответствует современным нормативам	6.3. Предлагаемое оснащение в целом соответствует современным нормативам	6.4. Предлагаемое оснащение в точности соответствует современным нормативам
7.	7.1. Симуляцию невозможно реализовать организационно и технологически. Она невыполнима	7.2. Симуляция может быть выполнена и реализована организационно и технологически с большим трудом	7.3. Симуляция организационно и технологически выполнима, реализуема	7.4. Симуляция абсолютно выполнима и реализуема как организационно, так и технологически
8.	8.1. Предложенный сценарий скучен, неоригинален, в нем отсутствует интересное зерно, изюминка, проблема для решения, некая провокация	8.2. В симуляции есть интересные моменты, но в целом отсутствует оригинальное зерно, изюминка, проблема, провокация	8.3. В симуляции имеется интересное зерно, изюминка, неожиданная проблема, провокация	8.4. Симуляция наполнена неожиданными интересными проблемами и «провокациями», захватывающими от начала и до конца

Среди экспертов РОСОМЕД было отобрано пять человек (А. А. Андреев, Р. Л. Буланов, И. И. Долгина, З. А. Зарипова, С. В. Ходус), которые не принимали участия в оценке референтного сценария и не были с ним ранее ознакомлены. Все рецензенты (Ж = 2, М = 3) имеют опыт руководящей и практической работы в симуляционном центре более 10 лет, являются практикующими клиницистами по специальности анестезиология-реаниматология (n = 4) и неонатология (n = 1). Экспертам были предоставлены все четыре документа без указания авторства и без уточнения, сколько было сценариев написано человеком, а сколько ИИ. Также была предложена таблица структурированной

оценки ШОСС с краткой инструкцией о ее применении. Экспертам было предложено проставить оценки по таблице ШОСС и определить, какие сценарии написаны человеком, а какие ИИ. Один из экспертов (С. В. Ходус) после завершения выполнения самостоятельной оценки без использования систем ИИ по собственной инициативе провел дополнительное исследование — составил промт [Врезка 2] и предложил системе ИИ DeepSeek на основе предложенной шкалы провести оценку четырех сценариев и определить авторство. Таким образом, в одном из случаев система ИИ оценивала собственный сценарий и пыталась определить авторство «ИИ/человек».

Промпт, данный системе ИИ DeepSeek для автоматизированной оценки четырех вариантов сценариев «Ты преподаватель в медицинском вузе, специалист симуляционного обучения в медицине, врач-анестезиолог-реаниматолог, работающий в России согласно всем нормативным актам и законам Российской Федерации. Твоя задача — провести экспертную оценку сценария симуляционного обучения. Критерии оценки сценария я представлю тебе в документе doc. После чего тебе будет представлен файл со сценарием. Проведи комплексную оценку и дай заключения по каждому критерию оценки сценария в соответствии с требованиями законодательства Российской Федерации. Представь комментарии по каждому из восьми параметров оценки. В конце проанализируй сценарий на оценку авторских прав и дай заключение о причастности искусственного интеллекта к написанию сценария. Думай глубже».

Результаты

Подробно результаты оценки сценариев представлены в таблице 2 ниже. Чтобы не перегружать настоящую статью табличными данными, оценки каждого из экспертов в отдельности не указаны, а приведены лишь средние значения (сумма баллов, деленная на 5 по количеству оценок) по каждому из пунктов шкалы ШОСС. Отдельной колонкой представлен разброс баллов и оценка, выставленная ИИ по каждому из сценариев, включая свой собственный.

Сценарий № 1, составленный генеративным ИИ Grok, получил самую низкую оценку как среди людей-экспертов, так и по мнению ИИ-оценщика DeepSeek (2,7 и 2,6 соответственно). Мнение экспертов в целом, совпадало, за исключением пункта реалистичность сценария, где было существенное расхождение мнений — с разбросом от одной крайней оценки до другой — от «1. Описанный в симуляции случай нереальный, в клинической практике так не бывает» (экспертом № 1) до «4. Описанный в симуляции случай реальный, случается в практике и требует тренинга и отработки» (тремя экспертами). Как видно и по ряду других пунктов таблицы, несмотря на структурность оценки и четкие формулировки каждого пункта, мнения экспертов разошлись — альфа Кронбаха составила 0,68, что свидетельствует об умеренной межэкспертной согласованности. Однако все, как эксперты, так и ИИ-оценщик, единодушно сошлись во мнении, что сценарий не был написан человеком, а создан компьютером.

Сценарий № 2, сгенерированный моделью ChatGPT, версия 5.0, получил как у экспертов, так и у ИИ-оценщика наиболее высокий балл, близкий к максимальному (3,4 и 3,5 соответственно). В этом сценарии эксперты проявили приемлемую межэкспертную

надежность — альфа Кронбаха составила 0,72. По сумме баллов сценарий занял первое место (3,4), получив также наивысшую оценку и ИИ-оценщиком (3,5). Четкие, конкретные, измеримые учебные цели; реалистичность сценария; соответствие оснащения нормативам, клиническим рекомендациям; выполнимость, реализуемость и, наконец, «интересный, проблемный сценарий» — по этим параметрам эксперты единодушно поставили оценки 3 и 4, т. е. «соответствует» или «превышает», другими словами, на «хорошо» и «отлично». Интересно, что ИИ-оценщик по тем же параметрам также поставил максимальную оценку, тем самым проявив солидарность с позицией людей. Примечательно и другое: не только трое из пяти специалистов признали автором данного сценария человека, но и ИИ-оценщик идентифицировал его как человеческое произведение, однако «созданное при поддержке чатбота».

Сценарий № 3, написанный авторами-людьми, получил вторую по высоте оценку со средним баллом 2,9, т. е. чуть ближе к критерию «соответствует ожиданиям». Недостаточно четко, конкретно и измеримо, по мнению экспертов (а также ИИ-оценщика), были прописаны учебные цели, тогда как по мнению специалистов и компьютера параметры «Соответствие оснащения нормативам, клиническим рекомендациям» и «Выполнимость, реализуемость» были сформулированы на «хорошо» и «отлично». Хотя большинство склонялось к человеческому авторству сценария, один из экспертов счел его творением чатбота.

И, наконец, четвертый сценарий, написанный генеративной системой DeepSeek, по оценке людей, занял третье, предпоследнее место. Примечательно, что при этом не только люди показали низкую межэкспертную

согласованность, но и позиция ИИ-оценщика существенно отличалась от человеческой. По параметрам «Соответствие УЦ уровню обучаемых» и «Описание приобретаемых навыков» различными экспертами была выставлена оценка от 1 до 4 — разброс мнений был максимальным. Средняя оценка, выставленная экспертами, составила 2,6 балла, тогда как ИИ-оценщик поставил документу оценку 3,8 — наиболее высокий балл из всех оценок, как человеческих,

так и компьютерных. Заслуживает внимание примечательный факт «компьютерного субъективизма» — ведь в данном случае именно DeepSeek был как автором, так и оценщиком данного сценария. Каким образом эта «предвзятость» возникла, судить сложно, ведь генерация сценария и его оценка производились с аккаунтов разных пользователей, из несвязанных между собой сетей и с компьютеров, которые разделяло расстояние более 7500 километров.

Таблица 2

Оценки сценариев по шкале ШОСС

1. Создан ИИ (Grok)		Среднее	Разброс	ИИ
1.	Соответствие УЦ уровню обучаемых	3,2	2–4	2
2.	Четкие, конкретные, измеримые УЦ	2,4	1–3	2
3.	Описание приобретаемых навыков	2,2	2	2
4.	Реалистичность сценария	2,8	1–4	4
5.	Соответствие навыков нормативам, клин. рекомендациям	2,2	1–2	1
6.	Соответствие оснащения нормативам, клин. рекомендациям	2,6	2–3	2
7.	Выполнимость, реализуемость	3	2–4	4
8.	Интересный, проблемный сценарий	3	2–4	4
9.	Средний балл	2,7		2,6
10.	Определение авторства			

2. Создан ИИ (ChatGPT)		Среднее	Разброс	ИИ
1.	Соответствие УЦ уровню обучаемых	3,2	2–4	2
2.	Четкие, конкретные, измеримые УЦ	3,6	3–4	4
3.	Описание приобретаемых навыков	2,8	2–4	4
4.	Реалистичность сценария	3,6	3–4	4
5.	Соответствие навыков нормативам, клин. рекомендациям	3	2–4	2
6.	Соответствие оснащения нормативам, клин. рекомендациям	3,6	3–4	4
7.	Выполнимость, реализуемость	3,6	3–4	4
8.	Интересный, проблемный сценарий	3,4	3–4	4
9.	Средний балл	3,4		3,5
10.	Определение авторства			

3. Создан человеком		Среднее	Разброс	ИИ
1.	Соответствие УЦ уровню обучаемых	3,2	2–4	3
2.	Четкие, конкретные, измеримые УЦ	1,8	1–4	2
3.	Описание приобретаемых навыков	2,2	2–3	2
4.	Реалистичность сценария	3	1–4	4
5.	Соответствие навыков нормативам, клин. рекомендациям	2,8	2–4	2
6.	Соответствие оснащения нормативам, клин. рекомендациям	3,4	3–4	3
7.	Выполнимость, реализуемость	3,4	3–4	4
8.	Интересный, проблемный сценарий	3	2–4	3
9.	Средний балл	2,9		2,9
10.	Определение авторства			

4. Создан ИИ (DeepSeek)		Среднее	Разброс	ИИ
1.	Соответствие УЦ уровню обучаемых	2	1–4	4
2.	Четкие, конкретные, измеримые УЦ	1,2	1–3	3
3.	Описание приобретаемых навыков	1,6	1–4	4
4.	Реалистичность сценария	3,2	2–4	4
5.	Соответствие навыков нормативам, клин. рекомендациям	3,2	2–4	3
6.	Соответствие оснащения нормативам, клин. рекомендациям	3	3–4	4
7.	Выполнимость, реализуемость	3,2	3–4	4
8.	Интересный, проблемный сценарий	3,2	3–4	4
9.	Средний балл	2,6		3,8
10.	Определение авторства			

Обсуждение

Таким образом, генеративные LLM-системы, представленные тремя платформами (Grok, ChatGPT, DeepSeek), показали принципиальную способность создавать развернутые клинические симуляционные сценарии по промпту, включающему четкие требования, такие как образовательные цели, структура, требования к нормативной базе и дебрифингу. Проведенный анализ подтверждает потенциальную применимость ИИ как инструмента поддержки разработки сценариев в симуляционном обучении. Наше исследование подтверждает возможность и даже целесообразность их использования на начальном этапе создания клинических сценариев в качестве инструмента для их первичной, черновой генерации с последующим редактированием документа человеком.

Экспертная оценка по шкале ШОСС выявила выраженную межмодельную вариативность качества: сценарий ChatGPT получил наивысший средний балл (3,4), превзойдя человеческий сценарий (2,9); сценарии Grok (2,7) и DeepSeek (2,6) оказались ниже человеческого по средним оценкам. То есть само по себе применение системы ИИ не гарантирует одинакового уровня результата: итоговый сценарий существенно зависит от конкретной модели и, вероятно, от параметров промпта и контекста применения.

Часть экспертов ошибочно принимала ИИ-сценарии за человеческие, и наоборот. Сценарий ChatGPT трое из пяти рецензентов идентифицировали как созданный человеком; при этом человеческий сценарий один эксперт счел произведением компьютера. Это указывает как на снижение «очевидности» машинного текста при высоком качестве генерации, так и на ограниченную валидность интуитивной атрибуции авторства в условиях слепого сравнения. Потенциал генеративного ИИ в создании качественных образовательных материалов для симуляционного обучения является высоким и недооцененным. Тот факт, что сценарий от ChatGPT получил оценку, сопоставимую с работой эксперта-человека, а по некоторым формализуемым критериям (четкость целей) превзошел ее, говорит о зрелости данной технологии.

Даже при структурированной шкале ШОСС согласованность рецензентов оказалась умеренной (максимум альфы Кронбаха достиг лишь в одном сценарии умеренной величины 0,72), а по ряду критериев разброс оценок достигал крайних значений. Это подчеркивает сложность и субъективность оценки симуляционных сценариев и необходимость дальнейшей отработки и уточнения шкал, более развернутой формулировки оценки, а также обучения экспертов.

ИИ-оценщик (DeepSeek), использованный для автоматизированной экспертизы по шкале ШОСС, в большинстве случаев продемонстрировал качество оценивания, сопоставимое с экспертами-людьми. При этом особого внимания заслуживает то, что сопоставимость проявлялась не во всех ситуациях: при самооценке сценария возникла выраженная модельная предвзятость (3,8 у ИИ против 2,6 у людей). Таким образом, несмотря на потенциальную привлекательность автоматизации ИИ-оценивания, этот процесс требует внешней валидации и контроля.

В дополнение к объективным выводам исследования хотелось бы представить сопутствующие наблюдения. Одним из существенных, едва ли не критическим недостатком авторства систем ИИ является необходимость проверки всех фактических данных — от точности результатов расчетов до корректности референтных ссылок — как литературных источников, так и клинической документации. Именно здесь, по наблюдениям авторов, наиболее ярко проявляется творческий, генеративный характер ИИ, который проявляется в ложных, выдуманных данных, так называемых галлюцинациях ИИ.

Другим практически важным аспектом на сегодняшний день является неспособность чат-ботов точно генерировать медицинские изображения, начиная с кривых (ЭКГ, ЭЭГ, плетизмограмм) и заканчивая данными рентгенологических исследований, КТ, МРТ, УЗИ. Если подобные файлы необходимы для реалистичности течения симуляционного сценария, то сейчас их может предоставить только автор-человек, облада-

ющий доступом к реальным базам данных медицинских изображений.

Выводы

Согласно данным, полученным в результате настоящего слепого сравнительного пилотного исследования, можно сделать следующие выводы:

- Существующие в настоящее время (ноябрь 2025 г.) системы генеративного искусственного интеллекта принципиально способны создавать симуляционные клинические сценарии и могут выступать эффективным инструментом ускорения и предварительной черновой разработки сценариев, особенно при четко заданных образовательных целях и нормативных ограничениях.
- Предложенный проект Шкалы оценки симуляционного сценария (ШОСС) может служить основой для структурированного рецензирования подобных организационно-методических документов. Однако высокая несогласованность экспертных выводов говорит о необходимости ее дальнейшего совершенствования, а также более тщательного предварительного инструктажа экспертов.
- Качество сценариев, разработанных системами ИИ и людьми, в целом сопоставимо по всем параметрам структурированной оценки. Требуется обязательная последующая экспертная правка, так как уровень сценариев отличается между платформами. ИИ пока не должен рассматриваться как автономная система, а может использоваться в качестве средства повышения эффективности и масштабируемости деятельности преподавателя.
- Системы ИИ, оценивающие сценарии по тем же структурированным параметрам, что и люди, как правило, дают сходную оценку. Однако использование ИИ для автоматической оценки сценариев следует расценивать лишь как вспомогательное средство; полностью заменять человеческое рецензирование преждевременно.
- Люди-эксперты не могут достоверно определить, кто является автором сценария — человек или ИИ. При высоком качестве текста ИИ-генерация

может быть неотличима от человеческой, что требует обсуждения этических и правовых аспектов в симуляционном сообществе.

- При оценке ИИ-системой различных сценариев обнаружилась предвзятость при самооценке. Несмотря на сопоставимые с человеческими результаты, в случае оценки собственного сценария ИИ-система дала оценку почти в полтора раза выше, чем люди.

Вклад авторов

Идея, дизайн и организация исследования, создание шкалы ШОСС, составление промпта, ИИ-генерация сценариев — М. Д. Горшков; проведение экспертной оценки сценариев по шкале ШОСС — А. А. Андреев, Р. Л. Буланов, И. И. Долгина, З. А. Зарипова, С. В. Ходус; составление промпта и ИИ-оценка сценариев — С. В. Ходус; текст — М. Д. Горшков, З. А. Зарипова, И. И. Долгина; составление исходного сценария — Н. Г. Косцова, В. Р. Непершина, О. В. Голубева; проведение «Лаборатории симуляции» и составление для нее таблицы оценки — Д. М. Грибков, Л. Б. Шубина.

ЛИТЕРАТУРА

1. Al-Elq A. H. Simulation-based medical teaching and learning. J Family Community Med. 2010 Jan; Vol. 17 (1). P. 35–40. DOI: 10.4103/1319-1683.68787. PMID: 22022669; PMCID: PMC3195067
2. Issenberg S. B., McGaghie W. C., Petrusa E. R., Lee Gordon D., Scalese R. J. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review // Med Teach. 2005 Jan. Vol. 27 (1). P. 10–28. DOI: 10.1080/01421590500046924. PMID: 16147767.
3. Kasneci E., Sessler K., Küchemann S., Bannert M., Dementieva D., Fischer F., Gasser U., Groh G., Günemann S., Hüllermeier E., Krusche S., Kutyniok G., Michaeli T., Nerdel C., Pfeffer J., Poquet O., Michael Sailer M., Schmidt A., Seidel T., Stadler M., Kasneci G. ChatGPT for good? On opportunities and challenges of large language models for education // Learning and Individual Differences. 2023. Vol. 103. № 102274. DOI: 10.1016/j.lindif.2023.102274